ACS NANO

www.acsnano.org

ARTICLE

AI-Based Prediction of Protein Corona Composition on DNA Nanostructures

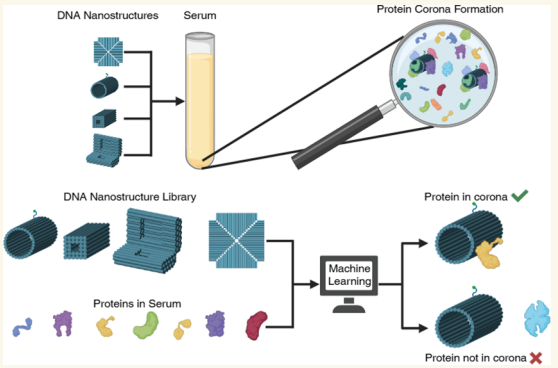
Jared Huzar,[†] Roxana Coreas,[†] Markita P. Landry, and Grigory Tikhomirov*

Cite This: https://doi.org/10.1021/acsnano.4c12259

Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: DNA nanotechnology has emerged as a powerful approach to engineering biophysical tools, therapeutics, and diagnostics because it enables the construction of designer nanoscale structures with high programmability. Based on DNA base pairing rules, nanostructure size, shape, surface functionality, and structural reconfiguration can be programmed with a degree of spatial, temporal, and energetic precision that is difficult to achieve with other methods. However, the properties and structure of DNA constructs are greatly altered *in vivo* due to spontaneous protein adsorption from biofluids. These adsorbed proteins, referred to as the protein corona, remain challenging to control or predict, and subsequently, their functionality and fate *in vivo* are difficult to engineer. To address these challenges, we prepared a library of diverse DNA nanostructures and investigated the relationship between their design features and the composition of their protein corona. We identified protein characteristics important for their adsorption to DNA nanostructures and developed a machine-learning model that predicts which proteins will be enriched on a DNA nanostructure based on the DNA structures' design features and protein properties. Our work will help to understand and program the function of DNA nanostructures *in vivo* for biophysical and biomedical applications.



KEYWORDS: *DNA nanotechnology, DNA origami, protein corona, biofouling, mass spectrometry, artificial intelligence (AI), machine learning*

Simple base pairing rules (A-T & G-C) have enabled the engineering of very complex, nanometer-precise DNA structures. Nanostructures consisting of 10,000 different DNA strands and reaching gigadaltons sizes with programmable behavior have been constructed.[1−4] These nanostructures can be designed to have precise arrangements of targeting ligands and dynamic reconfigurations with programmable kinetics.[5−7] DNA nanotechnology is now emerging as a versatile toolkit to study and alter biological processes.[8,9] As sensors, DNA constructs have been developed that can sense piconewton scale forces,[9] changes in temperature,[10] acidity,[11] and analyte presence.[12] As medicines, DNA nanostructures can sequester and on-demand release cargos[13,14] including functional nucleic acids (DNAzymes,[15] siRNA,[16] gene-encoding DNAs,[17] etc.) and proteins.[18] By modulating the size, shape, or addition of chemical moieties to nanostructures, DNA nanostructures can be programmed to reconfigure,[14,19] to target specific tissues,[20−22] and to be preferentially internalized by certain cell types.[23,24] However, all this powerful programmability of DNA nanotechnology is impacted by protein adsorption when nanostructures are immersed in biological environments. Analogous to protein adsorption that occurs on other materials, the protein corona changes DNA nanostructures' capabilities *in vivo*.

When DNA nanostructures are introduced to biological fluids (plasma, serum, etc.), they are spontaneously covered by a multitude of biomolecules forming the biomolecular corona.[20,25] The proteins that comprise the biomolecular corona drastically change how nanomaterials interact with biosystems[26] by altering nanoparticle size,[27] shape,[28] and physicochemical properties.[29] As the outermost entity, the biomolecular corona provides nanomaterials with a surrogate biological identity that can have unintended effects on nanostructure uptake, biodistribution, and immunogenicity.[26,30,31] Despite the biomolecular corona being a well-documented phenomenon across nanobiotechnology, the factors underlying biomolecular corona formation, especially

Received: September 2, 2024
Revised: December 21, 2024
Accepted: December 27, 2024

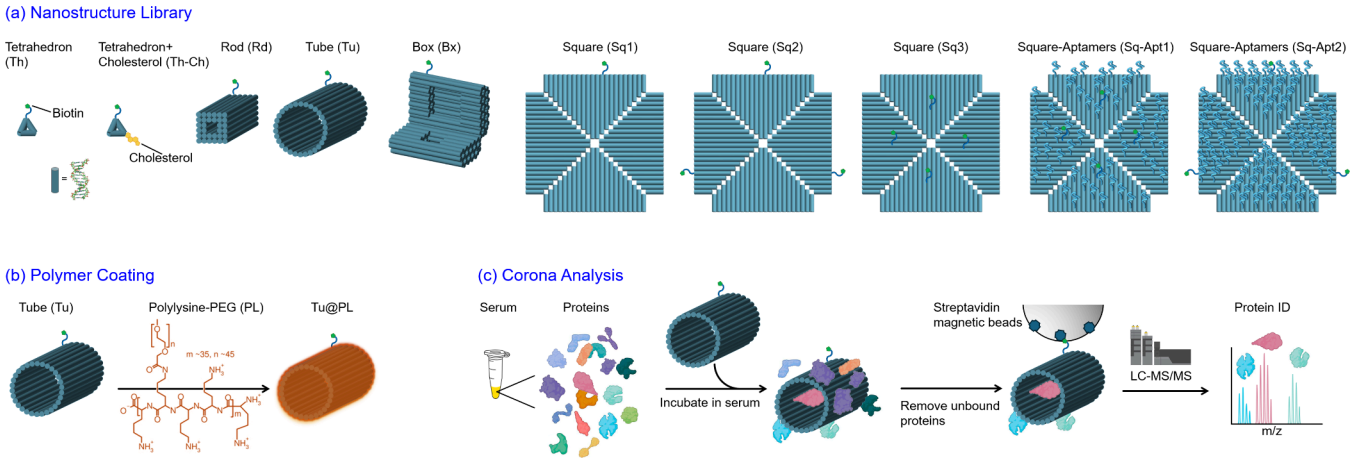


Figure 1. DNA nanostructures and protein corona analysis approach. (a) DNA nanostructures used in this study. (b) Schematic of nanostructures coating with PLL-*g*-PEG5K cationic polymer. (c) Schematic of protein corona analysis by magnetic bead separation and liquid chromatography tandem mass spectrometry.

for DNA nanostructures, remain insufficiently understood, limiting the applications of DNA nanotechnologies in biological fluids. Machine learning models have recently been used to elucidate the factors governing protein absorption on inorganic nanoparticles, liposomes, and carbon nanotubes.[32−34] Machine learning models have also been implemented to predict binding interactions between proteins and short nucleic acids,[35] however, *in silico* methods that accurately predict the interactions between biomolecules and DNA nanostructures, both *in vitro* and *in vivo* have yet to be developed and require extensive data sets built upon fundamental research. We sought to develop an interpretable machine learning classifier that can accurately predict which proteins will be found in the biomolecular coronas of DNA nanostructures.

To this end, we designed, synthesized, and characterized an array of DNA nanostructures with diverse design features including sizes, shapes, charges, and surface modifications like aptamers and cholesterol. We also synthesized several DNA nanostructures coated with oligolysine, a common polycationic polymer used to enhance cellular uptake and stability of DNA in biological environments.[36] With this library of nanostructures, we quantitatively measured the abundance of proteins adsorbed to the DNA nanostructures in human serum using gel electrophoresis shift assays and ultrahigh performance liquid chromatography tandem mass spectrometry (UHPLC-MS/MS). We observed differences along all DNA nanostructure design feature axes, i.e. certain proteins preferably adsorbed onto certain nanostructures.

With this rich data set, we developed an explainable machine learning model that can, based on basic DNA nanostructure and protein features, predict with 92% accuracy whether a protein will be present in the biomolecular corona. Thus, the protein corona can be predicted and even engineered with well-established DNA nanostructure design approaches. We leveraged this model to quantitatively probe relationships between size, shape, surface charge, and other features of DNA nanostructures and corona proteins. Thus, we gained insights into the factors governing protein adsorption and the biological pathways likely influenced by the adsorbed proteins. These findings will help guide the design of DNA nanostructures for biophysical and biomedical applications that are subject to biomolecular corona formation.

RESULTS AND DISCUSSION

Design Characterization of DNA Nanostructures and Protein Corona Analysis. We sought to elucidate the effect of DNA nanostructure (1) size, (2) shape, (3) surface functionalization, and (4) surface charge (Figure S1) on protein corona composition. These features are known to influence cellular uptake and other biological functions of DNA nanostructures.[20,23,37,38] To achieve this, we synthesized 17 DNA nanostructures (Figure 1a): (1) a DNA tetrahedron with 20mer duplexes per edge (Th),[20] (2) the same tetrahedron functionalized with a single cholesterol (Th-Ch), (3) an origami box (Bx),[39] (4) square origami tiles (Sq)[4] with differing numbers and positions of aptamers and biotins, (5) a hollow origami tube (Tu),[1] and (6) a 32 helix origami rod (Rd);[40] many of these structures were also synthesized with a polycationic PLL−PEG coating (@PL), a modification reported to influence stability, cellular uptake, protein corona formation, and surface charge[25,36] (e.g., tube in Figure 1b). Overall, these structures surveyed a wide design space over the various biologically relevant parameters (Figure 1a and S1).

Nanostructures were incubated in pooled human serum and they, along with their coronas, were purified to identify corona composition (Figure 1c). We performed gel electrophoretic analysis and observed that the proteins bound to DNA nanostructures were distinct from those bound to the magnetic bead in serum, our negative control, suggesting successful capture of our nanostructures and their coronas (Figure S2a). In addition, we observed a substantial difference in corona composition between the cholesterol-modified and non-cholesterol-modified tetrahedron (Figure S2b), corroborating previously reported results.[20] Having thus validated this protocol, we expanded this corona extraction technique to test several nanostructures at biologically relevant concentrations (10 pM − 100 nM) that are known to modulate cell inflammatory responses.[41,42] For higher-throughput and quantitative determination of corona composition, we performed SDS-PAGE separation followed by UHPLC-MS/MS. Across all DNA nanostructure coronas, UHPLC-MS/MS identified 575 proteins that showed differential abundance in the corona when compared to their prevalence in serum. Enriched corona proteins, ($\log_2$(fold change) > 0), were identified as proteins that were more abundant in the

https://doi.org/10.1021/acsnano.4c12259
ACS Nano XXXX, XXX, XXX−XXX

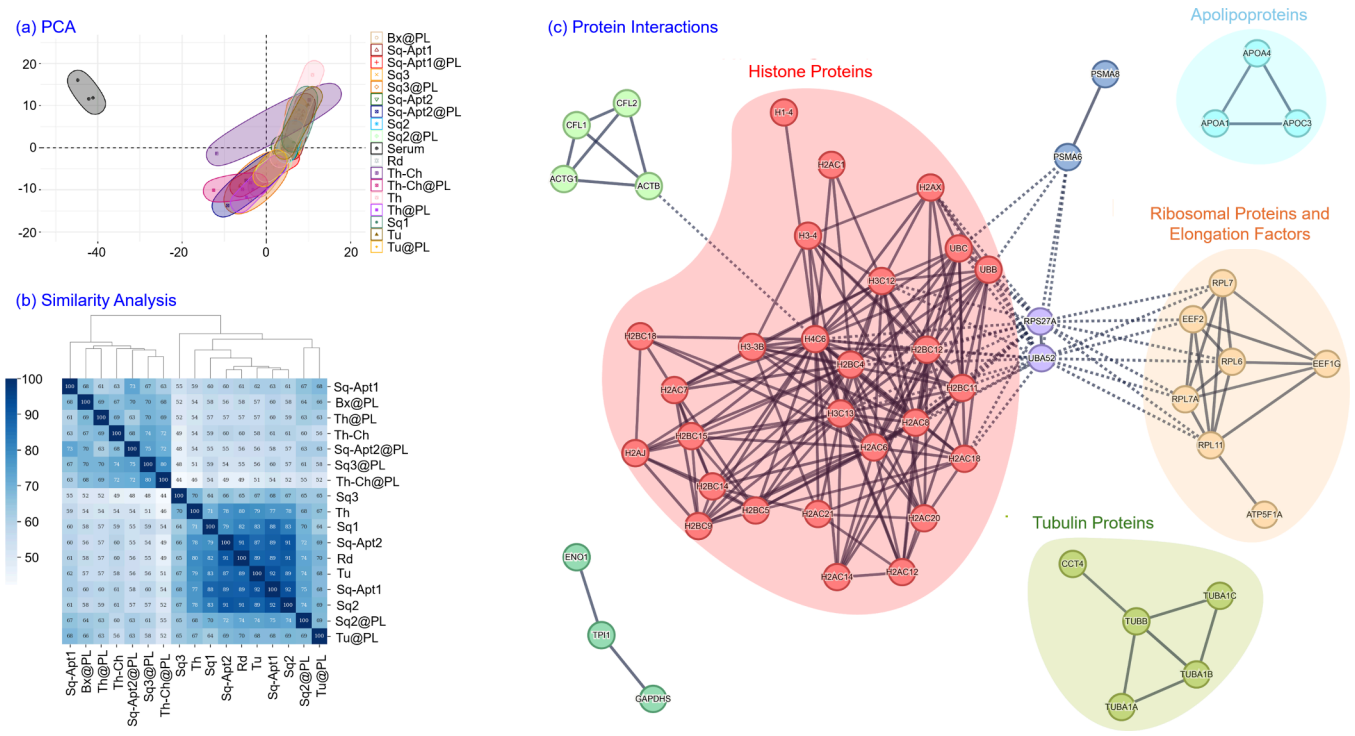


Figure 2. Protein corona analysis for polymer-coated and uncoated DNA nanostructures. (a) PCA of the protein coronas of all nanostructures and serum, with magnetic bead subtracted. PC1 and PC2 accounted for 37.3% (25.6% and 11.7%, respectively) of the total variance in the data. (b) Similarity heatmap plot with hierarchical clustering of the protein corona compositions for all nanostructures. (c) Interaction networks between proteins that are universally present across all nanostructures. Proteins with no interactions were removed.

nanostructure corona than in serum. Conversely, proteins that were more abundant in the controls, ($\log_2$(fold change) < 0), possessed minimal to no affinity to the nanostructures. The magnitude of change in protein prevalence for each nanostructure relative to the serum controls is shown in Figure S3; volcano plots for each nanostructure show the statistical significance between relative abundances (-$\log_{10}$(p-value)) versus the magnitude of change ($\log_2$(fold change)) of protein relative abundances (Figure S4). Within our data set, there were unique proteins that were present on the nanostructures but were not identified in serum nor on the magnetic bead, likely due to their low abundance in serum (Figure S5). Of these uniquely present proteins, several were involved in binding: Bx@PL enriched proteins involved in purine nucleotide and guanosine diphosphate binding; Sq-Apt1@PL and Sq3@PL enriched proteins associated in membrane adhesion, intracellular transport, and vesicle-mediated transport; Th-Ch@PL, Sq1, Rd, Sq3@PL, Sq-Apt1@PL and Tu enriched proteins involved with membrane docking; Th@PL enriched proteins that interact with biomolecules within the extracellular space and/or exosomes. Moreover, proteins associated with the positive regulation of early endosome to late endosome transport were enriched on Sq-Apt1, Sq-Apt1@PL, Sq3@PL, Sq-Apt2@PL, Rd, Th-Ch@PL, Sq1, and Tu. Several nanostructures enriched unique proteins associated with the positive regulation and/or activation of immunological processes; most notably, Sq3@PL enriched proteins involved in antimicrobial humoral responses, fibrinolysis, integrin activation, and regulation of ERK1 and ERK2 cascades; Sq-Apt1@PL enriched autocrine signaling proteins; molecular chaperones that fold stress-denatured proteins were enriched on Sq-Apt1, Sq-Apt2@PL, and Th; proteins associated with positive regulation of establishment of T cell polarity were enriched on Sq2; and proteins involved in wound responses and healing were enriched on Th-Ch@PL. Contrastingly, proteins involved in the negative regulation of immunological processes, including complement activation and regulation of extrinsic apoptotic signaling via death domain receptors, were enriched on Sq-Apt1@PL and Sq3@PL.

The magnitude of change in protein prevalence for each nanostructure relative to serum, after subtracting the protein's enrichment on the magnetic beads (MB), is shown in Figure S6. Recognizing these distinct protein preferences for nanostructures, we aimed to get deeper insights into these interactions. To do so we performed two different binary classifications of the individual proteins for each nanostructure: proteins that composed the nanostructure corona included those that were calculated as enriched or depleted, as well as those that were uniquely present on the corona but not in the sera controls. This threshold was decided as we expect all proteins present in the biomolecular corona will affect the nanostructures' physicochemical and pharmacokinetic properties, making it important to predict and understand the entirety of the corona's composition. However, we anticipate exploring enriched proteins will produce a model that is independent of protein concentration in the serum, making it more generalizable and even combinable with data from other biological fluids. Herein, we use the term "enriched" to describe proteins found in higher levels in the corona than in serum, "depleted" to represent proteins found in lower abundances on nanostructures relative to serum, and "present" to describe all proteins found in the corona irrespective of its abundance relative to the serum levels.

Differences and Similarities Across Origami. Considering the total corona composition of each nanostructure, we

https://doi.org/10.1021/acsnano.4c12259
ACS Nano XXXX, XXX, XXX−XXX

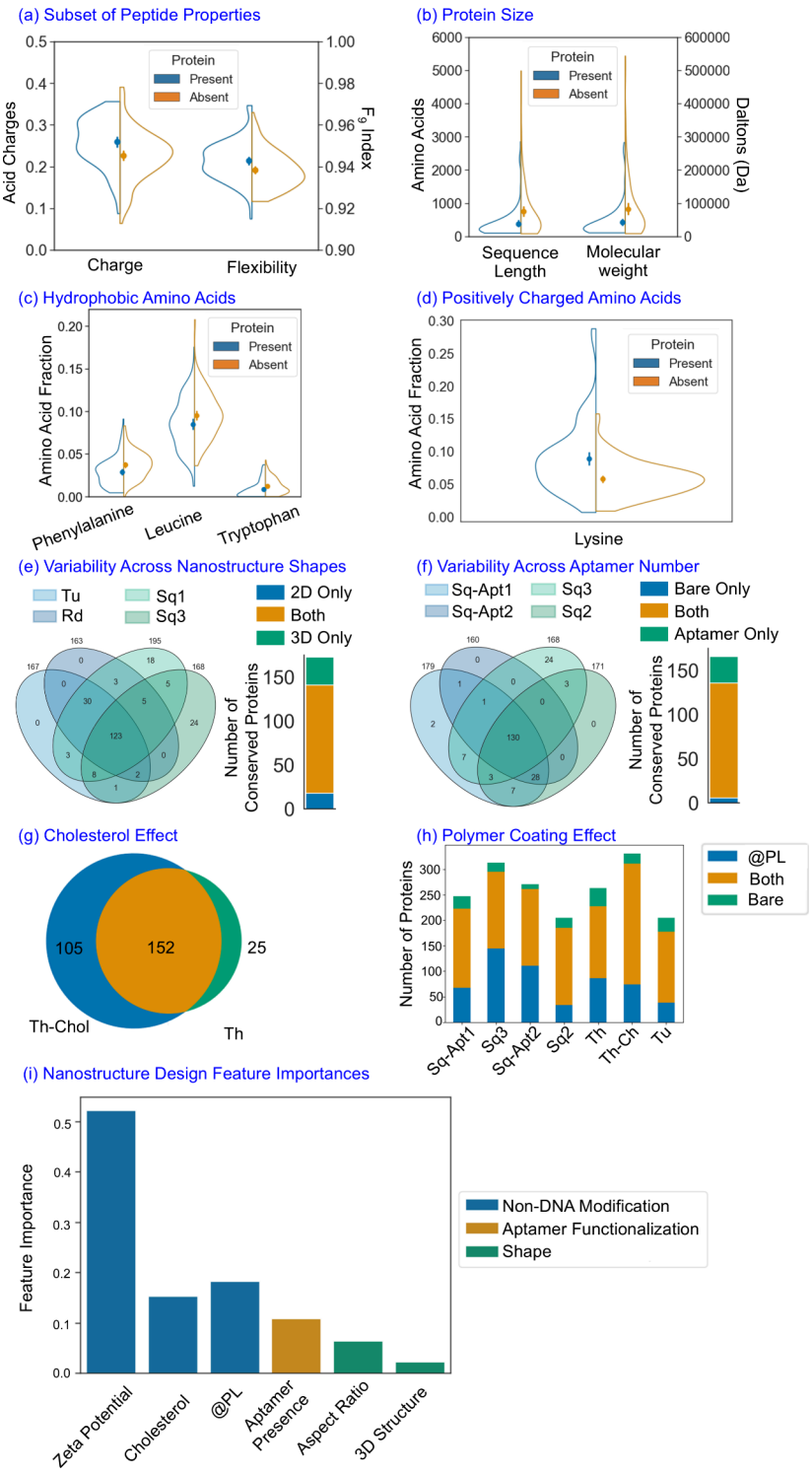


Figure 3. Effect of protein and DNA nanostructure properties on corona content. (a-d) Subgroups of protein properties that are significantly different between universally present and universally absent proteins from the corona. (a) Peptide charge and flexibility. (b) Peptide size. (c) Amino acids with hydrophobic side chains. (d) Amino acids with positively charged side chains. (e) Effect of the nanostructure shape: number of proteins belonging to one, some, or all the Rd, Tu, Sq1, and Sq3 nanostructures' coronas. Blue samples are 2D square sheets and green samples are elongated more rigid 3D shapes. (f) Effect of the number and position of aptamers or biotin tags on the same square origami: number of proteins belonging to one, some, or all the Sq-Apt2, Sq3, Sq2, Sq-Apt1 nanostructures' coronas. Blue samples are aptamer functionalized, and green samples are bare. (g) Effect of Cholesterol attachment: number of proteins in the corona of tetrahedron with cholesterol only, tetrahedron only, and both. (h) Effect of polymer coating: number of proteins present in the coronas of coated structures, their bare counterparts, and both. (i) Influence of various features on machine learning model prediction of proteins presence in nanostructures' coronas.

https://doi.org/10.1021/acsnano.4c12259
ACS Nano XXXX, XXX, XXX−XXX

observed all nanostructures differentiate themselves by their unique corona compositions relative to base serum composition (Figure 2a). We also observed minimal variation across the replicates of each nanostructure, but variation among the types of nanostructures and the serum-only control (Figure 2a). Interestingly, we find that corona content is not a simple reflection of each protein's relative abundance in the serum, rather that each nanostructure's corona has certain proteins that are enriched or depleted relative to their presence in human serum. To analyze this trend further, we considered the clustering of nanostructures based on their corona compositions. This revealed two distinct groups divided by the presence or absence of the polymer coating on the nanostructure. Among nanostructures with a polymer coating, there was a high degree, ~ 70%, of protein compositional similarity (Figure 2b). This homogeneity is even more pronounced among nanoparticles without a polymer coating (Figure 2b). Despite this polymer-driven clustering, 117 proteins—relative to 534 total analyzed proteins—were universally adsorbed across all nanostructures. These 117 universally adsorbed proteins exhibited a significant level of connectivity, referring to similarity regarding their endogenous biological roles (Figure 2c). Of the 117 proteins, several functional clusters emerged. Most predominantly, histone proteins represented a large fraction of these proteins, many of which are involved in the formation of nucleosomes, suggesting their propensity to interact with DNA.[43] In addition, other clusters were present consisting primarily of ribosomal proteins, tubulin proteins, and apolipoproteins among others. Finally, several immunoglobulin proteins were universally present. Immunoglobin proteins have been associated with the opsonization of nanoparticles,[44] and thus these results support the notion that the presence of foreign DNA in the form of nanostructures is likely to trigger their clearance. Notably, this opsonization phenomenon is likely to occur regardless of nanostructure size and shape. Both apolipoproteins and immunoglobins are commonly present in the coronas of other organic nanoparticles, like liposomes[45,46] and inorganic nanoparticles,[47] as well as these DNA nanostructures. As expected, albumin, a highly abundant, amphiphilic protein was found across most of the nanostructures' coronas, consistent with inorganic gold nanoparticles.[48] Interestingly, thrombospondin-1 (TSP1), which interacts with several cell adhesion receptors,[49] was preferentially absorbed on all nanostructures with the @PL coating, with the highest enrichment observed on Th@PL. Nanostructures coated with @PL also enriched plasma serine and protein z-dependent protease inhibitors; the highest enrichment of these modulatory serpins that regulate inflammatory responses[50] was observed on Box@PL. Other immunomodulatory proteins that were preferentially adsorbed onto the @PL-coated nanostructures include carboxypeptidase B2,[51] most abundant on Th@PL, alpha-2-HS-glycoprotein (fetuin-A) and its less abundant homologue fetuin-B,[52] most abundant on Th@PL and Sq3@PL, respectively, phospholipid transfer protein,[53] most abundant on Sq-Apt1@PL, lipopolysaccharide-binding protein,[54] most abundant on Sq3@PL, CD5 antigen-like (CD5L),[55] most abundant on Bx@PL, complement components C6, C9, and factor B,[56] most abundant on Th@PL, Sq2@PL, and Tu@PL, respectively. Contrastingly, ELAV-like protein 1, which suppresses inflammatory responses,[57] was preferentially adsorbed on the noncoated nanostructures, with the highest enrichment observed on Tu. Interalpha-trypsin inhibitor heavy chain 2 (ITIH2), which inhibits complement activation,[58] was also preferentially adsorbed by noncoated nanostructures, with the highest enrichment observed in Sq1. However, several immunogenic proteins were also preferentially absorbed on the noncoated nanostructures and not their @PL-coated counterparts, including elongation factor 2,[59] which was most abundant on Rd, vitronectin,[60] most abundant on Sq1, and complement C1q subcomponent subunit C,[56] most abundant on Sq3. Additionally, histone proteins H1, H2, and H3, involved in packaging DNA into chromatin and transcriptional activation,[61] were preferentially enriched on nanostructures lacking the @PL coating; Sq3 had the highest enrichment of histone proteins. Apolipoproteins A-I and B-100, the former having immunogenic properties and the latter potentially exhibiting such properties,[62,63] were enriched on several nanostructures, irrespective of @PL coating presence, yet Tu@PL and Th-Ch@PL had the highest enrichments, respectively.

Effect of Protein Properties. We anticipated the 117 universally present corona proteins would contain common physicochemical features that influence their adsorption to nanostructures irrespective of nanostructure polymer coating presence or absence. To examine this, we developed a database of functional, structural, and physicochemical properties of proteins to identify the protein features important in determining a protein's abundance in the corona (see the methods section for more information on the specific metrics). Overall, our database leveraged single amino acid level properties, secondary structure information, and functional information to encompass protein properties most likely to affect protein corona adsorption. Indeed, when compared with proteins that were not found in the corona of any nanostructures, clear trends emerge differentiating the universally adsorbed and universally absent proteins. A set of 30 protein physicochemical properties exhibited statistically significant distributions (p-value <0.01) (Table S2). Highlighting a subset of these 30, several interesting patterns emerged. We observed that proteins with greater flexibility, as defined by its F_9 index,[64] and a more positive charge exhibited an increased propensity for adsorption to the nanostructures (Figure 3a), which makes sense considering that DNA nanostructures are negatively charged and relatively rigid. A similar phenomenon of flexible proteins being enriched in a nanoparticle corona was observed with single-walled carbon nanotubes (SWCNTs),[33] a rigid, inorganic nanoparticle. Interestingly, larger proteins were less likely to be adsorbed (Figure 3b). It has been previously reported that the adsorption of small proteins onto nanostructures is an enthalpy-driven process, versus larger proteins that adsorb via an entropy-driven mechanism.[44] As such, we expect this protein size-dependent adsorption observation may be temperature-dependent, as well as nanostructure-dependent. For this reason, we performed all corona adsorption experiments at 37 °C to best mimic conditions experienced *in vivo*. We found that proteins' different amino acid compositions elicited different effects on the proteins' adsorption to nanostructures. For example, proteins with larger fractions of amino acids with hydrophobic side chains (phenylalanine, leucine, and tryptophan) were less likely to be adsorbed to hydrophilic DNA nanostructures (Figure 3c). However, positively charged lysine residues were associated with greater protein adsorption (Figure 3d).

https://doi.org/10.1021/acsnano.4c12259
ACS Nano XXXX, XXX, XXX−XXX

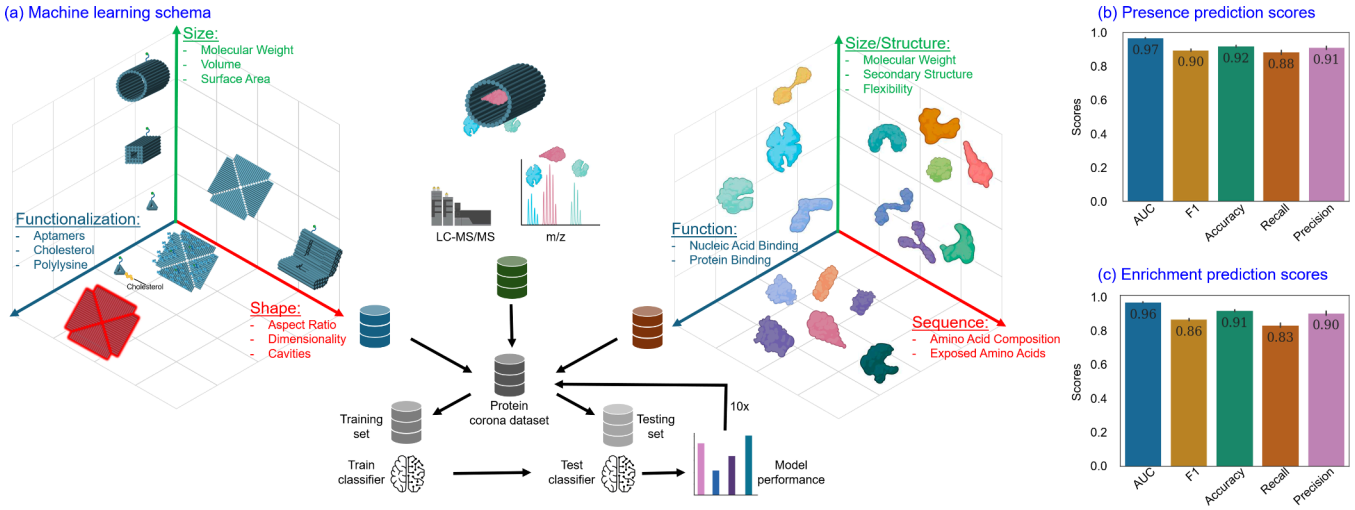


Figure 4. Machine learning model scheme and performance. (a) Schema for machine learning model and evaluation protocol. (b) Performance metrics of the XGBoost model classifying proteins as present or absent in the nanostructures' coronas across 10 splits. (c) Performance metrics of the XGBoost model classifying proteins as enriched or depleted in the nanostructures' coronas.

Effect of Nanostructure Design Properties. Using our library of nanostructures, we next probed the effects of nanostructure design features on corona composition by quantifying variation in protein corona composition between DNA constructs with similar and dissimilar designs. We first investigated the influence of nanostructure shape on the protein corona composition. Comparing Sq3 and Sq1 (two similar flat, square structures) to Rd and Tu (two elongated, 3D structures), we can consider the effect of aspect ratio and dimensionality (2D versus 3D) while keeping size and physicochemical properties mostly constant. If nanostructure shape has a large effect on protein corona composition, we would expect compositional similarity between the long, 3D structures (Rd and Tu) and between the flat, square structures (Sq3 and Sq1), but large compositional differences across the nanostructures with different shapes (Rd and Tu vs Sq1 and Sq3). However, we observed that many proteins are commonly found in all four structures, and there are few proteins conserved across either shape category that are not conserved across the other (Figure 3e). Of the proteins found in the corona of any of the four structures, 55% are commonly found in all 4 nanostructures' coronas. This demonstrates the marginal effect of DNA origami structure on protein corona composition. Consequently, we conclude DNA nanostructure shape has a limited role in protein adsorption for nanostructure shapes tested herein. Next, given the utility of aptamers in biological applications involving DNA nanostructures, we investigated the effect aptamer number and positioning had on corona formation on otherwise identical DNA nanostructures. We functionalized our nanostructures with the CoV2-RBD-1C aptamer, discovered by Song et al.[65] since this aptamer should have no specific affinity for proteins in the serum. To parse the effect of aptamers, we compared Sq2 and Sq3 to Sq-Apt2 and Sq-Apt1 (the aptamer functionalized equivalents of Sq2 and Sq3). If aptamer presence has a significant effect on corona composition, we would expect to see a large degree of similarity between Sq-Apt2 and Sq-Apt1, but an observable difference between nonfunctionalized (Sq2 and Sq3) and their aptamer-functionalized counterparts (Sq-Apt2 and Sq-Apt1). We observed that of proteins found in the corona of any of the four structures, most, 63%, are commonly found across all four structures, and no clear distinctions appear among the aptamer-modified structures to differentiate themselves from their nonfunctionalized counterparts (Figure 3f). We also glean insight into the effect of aptamer placement and multivalency, a property important in biological applications of nanostructures like cell labeling,[66] by comparing Sq-Apt1 to Sq-Apt2. Sq-Apt1 has all aptamers on a single face, while Sq-Apt2 has all aptamers on both faces of the square DNA nanostructure. We hypothesized a greater number of aptamers and more complete decoration (Sq-Apt2), could result in a reduction in protein adsorption due to steric hindrance, an effect previously observed with other polymers.[67] However, most proteins adsorbed on each structure are commonly found in both of their coronas, and there is minimal change in the diversity of proteins adsorbed on each (Figure 3f). These results emphasize the limited effect of aptamer presence, number, or position on the protein corona.

We next investigated the effects of cholesterol modification as well as the @PL polymer coating since non-DNA modifications are known to significantly affect corona composition. Comparing the tetrahedron with and without cholesterol, we observed that a cholesterol modification causes an almost entirely unique protein corona in addition to the typical corona of the DNA-only nanostructure. The Th-Ch nanostructure corona had 105 proteins found exclusively in the structure, in addition to 152 proteins found in both the Th-Ch corona and Th corona. Th only had 25 unique proteins in its corona (Figure 3g). In a similar manner, by comparing seven uncoated nanostructures to their polymer-coated counterparts, we were able to study the effect of the coating and change in surface charge independently of other variables (Figure 3h). Across the seven nanostructures, each of which was coated with the @PL polymer, most (up to 74%) of proteins adsorbed were present in both the polymer-coated nanostructure and uncoated nanostructure coronas, although generally more were also uniquely present to the polymer-coated nanostructures' coronas as compared to the bare structures.

To further elucidate the influence of nanostructure design features on corona composition, we compared the feature importance values obtained from our explainable AI model (explained in greater detail below). We observed that non-

https://doi.org/10.1021/acsnano.4c12259
ACS Nano XXXX, XXX, XXX−XXX

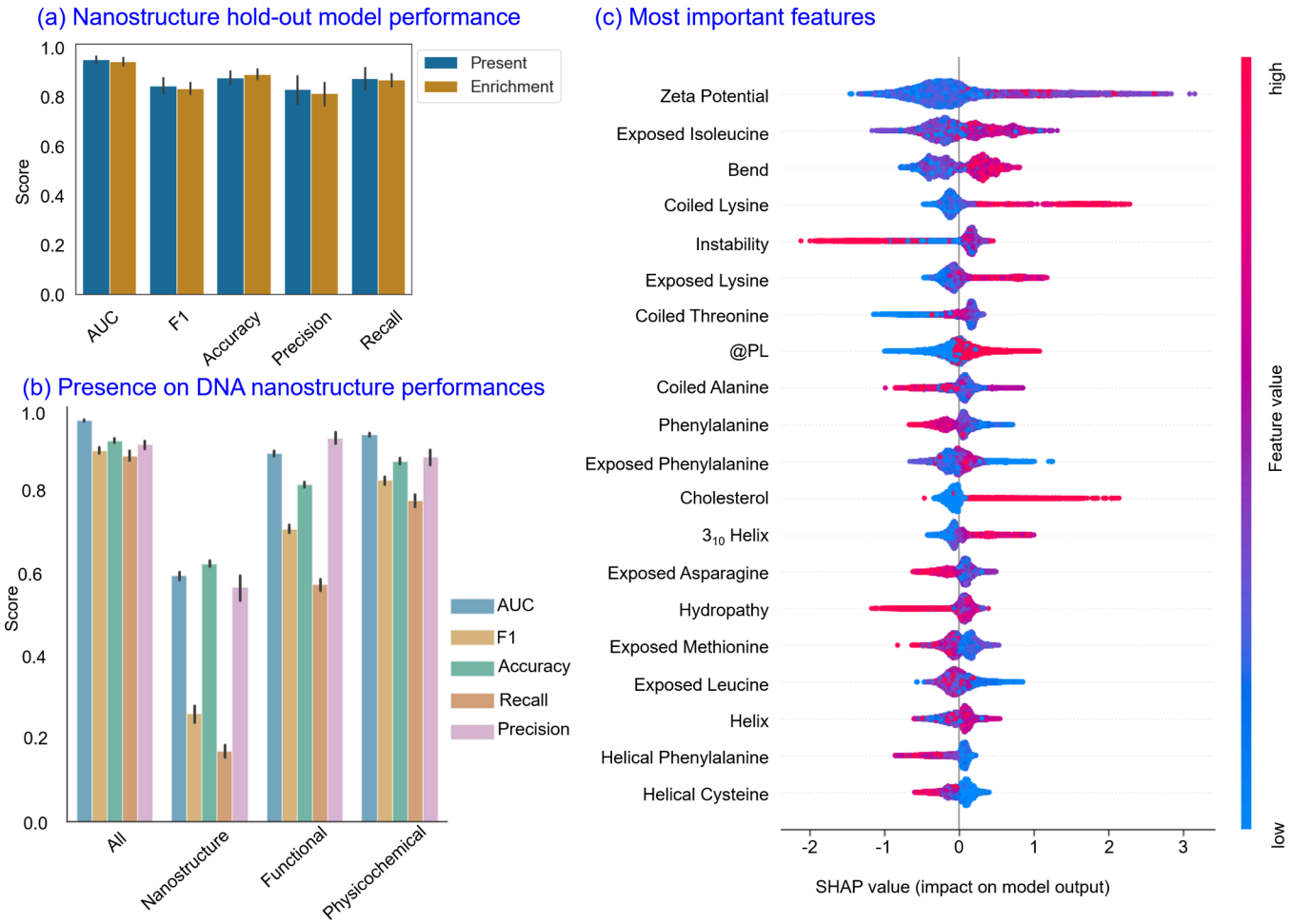


Figure 5. Model performance on "held-out" nanostructures and data subsets, and model feature importances. (a) Mean performances of the XGBoost model trained on all but one nanostructure and tested on the held-out nanostructure. (b) Mean performances of the XGBoost model in classifying proteins as present or absent when trained with different subsets of the data. (c) SHAP value plot of the 20 most important features for classifying proteins as present or absent.

DNA modifications are more influential than both aptamers and nanostructure shape (which are relatively similar in importance) (Figure 3i).

Developing a Machine Learning Model to Understand and Predict Nanostructure Protein Corona Formation. To better understand the factors influencing the differences in corona formation on DNA nanostructures, we sought to develop an explainable machine learning model. We elected to study two classification tasks: the prediction of enriched proteins and the prediction of all proteins present within a DNA nanostructure's protein corona. The two classification tasks both have their unique advantages. We expect the prediction of enriched proteins will create a more generalizable model with less of an effect from initial protein concentration in the serum. However, prediction of all proteins present is essential to understanding and engineering nanoparticle fate. To train and implement the model, we combined our assembled protein features database with nanostructure physicochemical properties (Tables S1,3) and with the UHPLC-MS/MS as a ground truth (Figure 4a). We chose to implement the XGBoost[68] algorithm as it is an implementation of gradient-boosting decision trees that maintains the interpretability of decision trees while offering advantages in speed and accuracy. Also, XGBoost has previously been demonstrated as an effective algorithm for protein corona classification.[69] To evaluate the validity of this approach, we tested two other ensemble methods that use sklearn:[70] (1) a Random Forest classifier[71] and (2) a Gradient boosting classifier.[72] XGBoost had superior performance in identifying both enriched and present proteins upon evaluation using the area under the receiver operating curve (AUC), accuracy, f1, precision, and recall (Figure S7). We subsequently chose to use this architecture for the remainder of the analyses.

Combining the data from all 17 nanostructures into one training/test data set, our XGBoost model achieved 0.97 AUC and 92% accuracy in classifying a protein as present or absent from the protein corona (Figure 4b). This demonstrates its high performance in identifying proteins likely to be adsorbed across an array of nanostructures. Thus, we validated our model as a useful tool allowing for the prediction of proteins found within the protein corona on nanostructures. Utilizing these findings, researchers can now preemptively predict, and account for, the proteins likely to be found on their DNA nanostructures, a first-of-its-kind tool to our knowledge. In addition, using the importance of each feature, this model can be informative about which nanostructure design characteristics can most influentially bias protein adsorption (Figure 3i). And finally, we can use this model to identify characteristics of adsorbed proteins that are generalizable across nanostructures.

https://doi.org/10.1021/acsnano.4c12259
ACS Nano XXXX, XXX, XXX−XXX

Using the same architecture, we classified proteins that are enriched (i.e., more abundant in the corona than in serum) on nanostructures. For this novel classification task, we found an XGBoost model was able to maintain a high level of performance, achieving an AUC of 0.96 and an accuracy of 91% (Figure 4c). Since this model effectively subtracts the serum concentrations of proteins, we expect insights derived from this explainable model can be generalizable to protein adsorption onto DNA nanostructures from other biological fluids. Thus, we successfully implemented and characterized two models with high power for predicting proteins likely to impact the nanostructures' physiological and pharmacokinetic properties. The models' high performances also validate them as tools from which we can gain insights into the factors governing protein adsorption on DNA nanostructures.

For applications, it is important to have an assessment of a prediction's reliability. For each binary classification of a protein as present or absent in the corona, the model predicts a probability for each outcome. We explored whether this probability could act as a measure of confidence for the predictions. To test this, we compared the distribution of protein in-corona probabilities for predictions that were true positives, false positives, false negatives, and true negatives in a test set. The model on average assigned higher in-corona probabilities for true positives than false positives, and lower in-corona probabilities for true negatives than false negatives (p-value <0.01) (Figure S8). This finding demonstrates that prediction probabilities can be used as measures of a prediction's reliability.

To validate our tool's potential for *a priori* prediction of protein corona composition on novel nanostructures, we trained our model on all but one nanostructure, and then tested the model's performance on the remaining "held-out" structure, a process we repeated 17 times until all structures were tested. In this manner, we explored the ability of our model to generalize to new nanostructures and determined whether it could be applied to help guide the design and implementation of novel nanostructures researchers wish to use for experiments in biological fluids. The model maintained high performance in classifying protein presence (0.95 AUC and 88% accuracy) and protein enrichment (0.94 AUC and 89% accuracy) (Figure 5a). Thus, we demonstrated our architecture's utility for the smart, informed design of new nanostructures.

Finally, we experimented with training a unique model for individual nanostructures. Despite this being a simpler classification task, these models performed much more poorly (Figure S9), likely because there is significantly less data to train on as compared to the bulk data set.

Important Protein and Nanostructure Features Governing Protein Adsorption. Since we validated the predictive power of our model, we next utilized its interpretable decision-making architecture to discover which features were most informative and predictive to the formation of the nanostructure protein corona. We tested over 600 features quantifying protein size, structure, amino acid composition, and functionality, and DNA nanostructure size, shape, and functionalization. For the models predicting protein enrichment and presence, we found that more than 150 out of over 600 tested features contributed to the classification of the proteins (Table S3). This result underscores the vast complexity of factors governing protein corona adsorption. To elucidate on a broader scale the properties identified as drivers of protein corona formation, we subset our data into 3 classes: (1) Protein sequence, structure, and physicochemical features, (2) Protein functional features, and (3) DNA nanostructure design features (Table S3). Training the data on each of these subsets alone saw reduced performance as compared to the original data set containing all this information. Protein sequence, structural, and physicochemical features were the most effective at predicting protein presence in the corona, followed by protein function, and last origami design (Figure 5b). Specifically, protein sequence, structural, and physicochemical features alone effectively predicted protein adsorption with 87% accuracy, protein function predicted protein adsorption with 81% accuracy, and origami design predicted protein adsorption with 62% accuracy. Since the protein features alone were not as effective as when combined with origami design features, we conclude that protein features governing corona adsorption are not generalizable across all nanostructures (i.e., different nanostructures preferably adsorb proteins with different properties). However, the moderate predictive power of protein sequence, structure, and physicochemical properties alone suggests that this feature class plays a large role in determining protein in-corona presence. DNA nanostructure design features were the weakest predictor indicating that it is feasible to build a general predictive tool for protein corona formation on a large variety of DNA nanostructures of relevance across diverse biomedical applications.

To interpret the importance values of different features, we calculated the Shadley Additive exPlanations (SHAP) values for each feature.[73] According to the SHAP values, two of the ten most important features were related to the modifications of DNA nanostructures: @PL coating and ζ-potential (Figure 5c). Low values of ζ-potential and the absence of the @PL coating are both associated with proteins not being present in the corona. The presence of a cholesterol-modified DNA strand is another influential design choice. Cholesterol had a particularly large impact on the model as its presence strongly and positively influenced whether a protein would be adsorbed. This finding further supports our conclusion that non-DNA modifications affect the protein corona composition on DNA nanostructures more than construct design considerations like size and shape.

Several insights regarding protein properties governing adsorption can be gleaned from the SHAP values. For example, some protein secondary structures, like 3_{10} helices and bends, promoted positive predictions (Figure 5c), indicating that proteins containing a large fraction of 3_{10} helices and bends in their secondary structures are more likely to be predicted as present in the corona. Different amino acids had different effects on a protein's likelihood to adsorb. For instance, proteins with a large fraction of exposed isoleucine residues corresponded with positive protein adsorption predictions, while proteins with a large fraction of exposed asparagine residues led to negative predictions of protein adsorption (Figure 5c).

We then explored feature importances from the models trained on subgroups of DNA nanostructures with similar properties, i.e. are certain protein features better for protein adsorption on @PL coated vs uncoated nanostructures. Comparing the SHAP values for models trained on data from different structures can provide further insights into which protein properties are more likely to lead to adsorption across nanostructures of different design axes. However, we

https://doi.org/10.1021/acsnano.4c12259
ACS Nano XXXX, XXX, XXX−XXX

had to consider the reduction in predictive power evident when we train models on individual nanostructures (Figure S9). So, to explore the effect of coating the nanostructures, we trained one model on all the nanostructures coated with @PL and separately trained another model on all nanostructures without a coating. We first validated these models and found that they maintained high levels of predictive power, with the models considering the uncoated and coated structures separately achieving accuracies of 94% and 90% respectively (Figure S10a, b). From the 20 most important protein and nanostructure features for each model, 7 protein properties (bend secondary structure, exposed isoleucine, coiled lysine, exposed asparagine, helical cysteine, protein instability, and phenylalanine) are commonly influential to both coated and uncoated nanostructure predictions (Figure S10c, d). Among these 7 features, each had similar relationships regarding feature effect on protein corona presence across both models, indicating there are some conserved principles governing adsorption to DNA nanostructures regardless of the polymer coating. However, many of the important features are not conserved across models, suggesting different protein properties affect adsorption differently across coated and uncoated structures. This can explain the corona composition differences we see across coated and uncoated structures (Figure 2b).

Overall, our results demonstrate that several protein features affect protein adsorption onto DNA nanostructures of varying sizes, shapes, and modifications differently; while several other features, like increased flexibility and decreased amino acid sequence length, promote practically universal adsorption onto the DNA nanostructures. These results suggest it is possible to bias protein corona composition with nanostructure engineering, albeit with incomplete control over the entire proteome. Our results also suggest that engineering the biofluid itself, or perhaps precoating nanostructures with specific proteins, could enable greater control over DNA nanostructure physicochemical identity for subsequent use in a range of biofluids.

Lastly, we sought to understand the role of protein function on corona composition. Protein functions were defined as the gene ontologies that are classified as molecular functions, which included ATP binding, actin binding, helicase activity, and many more. Specifically, for protein functions found in at least 23 of the 534 proteins identified across all analyzed trials, we calculated the enrichment score for each protein function. We define enrichment score as the fold difference in the number of different proteins with that particular function present, as compared to the number of proteins with said function expected by chance to be in the corona. We find that of all tested protein functions, several are significantly enriched and depleted in proteins found within the corona (p-value <0.05) (Figure 6a). These significantly enriched functions fall into two primary categories: (1) nucleic acid binding and (2) protein binding. As expected, DNA-binding is an enriched function, as well as RNA-binding, with a 1.7- and 1.2-fold increased diversity of corona proteins exhibiting DNA or RNA binding functions respectively, as opposed to the number expected by random. The most enriched functional group was the structural constituent of chromatin group, with a 2.2-fold enrichment of corona proteins being chromatin constituents. These proteins endogenously interact with DNA given they are part of the complex of DNA and proteins that make up chromatin, as such, it is reasonable that these protein functions would positively influence protein adsorption to nanostructures. The remaining functions comprise the second category:

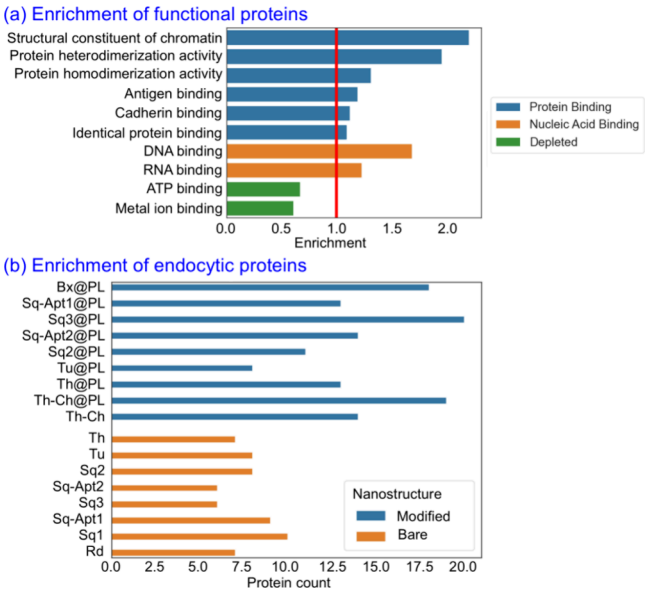


Figure 6. Differential adsorption of proteins with distinct functions in the corona. (a) Enrichment of functional protein families in nanostructure coronas. All enrichment and depletions are statistically significant (p-value <0.05). (b) The number of endocytosis-associated proteins across modified and bare DNA nanostructures.

protein binding, suggesting proteins with a propensity to bind with other proteins are more likely to be present in nanostructures' coronas. Protein heterodimerization activity exhibited a 2.0-fold enrichment, protein homodimerization activity a 1.3-fold enrichment, antigen binding a 1.2-fold enrichment, cadherin binding a 1.1-fold enrichment, and identical protein binding a 1.1-fold enrichment. We hypothesize that since the protein corona consists of several layers,[74] the outermost layer (soft corona) at any given time affects the interactions of the remaining biofluid proteins. Therefore, the ability of a protein to be present in the outermost layer of a corona may be dependent on its ability to bind to and interact with other proteins forming the innermost (hard corona) layers. Interestingly, there was a large depletion in metal ion binding and ATP binding proteins, with 0.6 and 0.7-fold depletions, respectively. This occurs possibly because metal ion binding proteins have a greater affinity to the metal ions in the nanostructure buffer than the nanostructure itself. ATP binding proteins were depleted likely due to steric hindrances preventing nanostructures from compatibly interacting with the binding domain of the ATP binding proteins. We conclude that, of all considered protein features, a protein's degree of nucleic acid binding and protein binding are the most influential in its presence or absence in the corona of DNA nanostructures.

Having demonstrated statistically significant enrichment/depletions of various functional protein families, we explored if proteins associated with different biological processes can be differentially adsorbed to the surface of DNA nanostructures intentionally. Specifically, we examined whether the protein corona could be engineered by changing the different design parameters of the nanostructures. As a proof of concept, we explored engineering the nanostructure protein corona with proteins involved in endocytosis (gene ontology group GO:0006897) in different nanostructures. We selected this proof-of-principle experiment because controlling and better

https://doi.org/10.1021/acsnano.4c12259
ACS Nano XXXX, XXX, XXX−XXX

understanding endocytosis is important for cellular delivery of therapeutics and biophysical tools such as DNA nanostructures. We found that design axes like DNA nanostructure size and shape yielded minimal changes in the diversity of unique endocytic proteins adsorbed to the nanostructure. However, we observed an almost doubling in the diversity of endocytic proteins in the corona of modified (cationic polymer or cholesterol functionalized) nanostructures versus their bare, unmodified counterparts (Figure 6b). Therefore, we hypothesize that cationic polymer coatings can enrich the protein corona for proteins associated with endocytosis. This effect is synergistic with the previously reported improvement in DNA nanostructure stability and ζ-potential increase when cationic polymers are used to coat nanostructures.[25,36] Our results suggest that introducing non-DNA modifications to the surfaces of DNA nanostructures may be able to modulate the composition of the protein corona across other protein functional domains.

CONCLUSION

Herein, we broadly surveyed the effect of DNA nanostructure design parameters on the composition of their protein coronas when incubated in human serum. We find that modulating structural properties of nanostructures (size, shape, etc.) can lead to differential adsorption of a minority of the overall proteins present in the corona, but that nanostructure design is largely less influential in driving protein corona composition than non-DNA nanostructure surface modification with polymers or cholesterol. Design features of the nanostructures only slightly bias the adsorption of different proteins, with at most 36% difference in protein corona composition between the two most dissimilar DNA-only nanostructures. Conversely, the addition of non-DNA modifications (cholesterol and cationic polymer coatings) leads to the most pronounced changes in coronas, with up to 52% difference in protein corona composition between the polymer-coated nanostructure relative to its uncoated counterpart. We hypothesize this non-DNA modification-driven increase in protein corona compositional diversity is due to the hydrophobicity of cholesterol and the cationic charge of the polymer attracting new classes of proteins to the corona, thereby adding to those already binding to hydrophilic, negatively charged DNA. Our work demonstrates the potential to engineer the nanostructure corona using both non-DNA modifications to the nanostructure and, to a lesser extent, by modifying the DNA nanostructure itself.

In DNA nanotechnology, the protein corona has been engineered to have implications on nanostructures' biological activities like cellular uptake[25] and endosomal escape.[31] To further promote corona engineering, we developed two explainable machine learning models that predict whether a protein will be present/absent or enriched/depleted from a given nanostructure's corona. These models are first-of-their-kind tools enabling the accurate prediction of the protein corona on any given DNA nanostructure with up to 92% accuracy. In addition, by utilizing an explainable algorithm, the models offer valuable insights into the factors governing the adsorption of proteins, both in the case of proteins that ubiquitously bind all DNA nanostructures versus proteins that are unique to certain nanostructure constructs. Therefore, we envision our model will enable researchers to further reclaim the programmability of DNA nanostructures that is typically lost with spontaneous protein corona formation. With model-based predictability of nanostructure protein corona composition, researchers can account for spontaneous protein adsorption prior to experimentation. Furthermore, our approach can support the design of nanostructures with designer coronas, an effort toward which has already begun.[75] Utilizing the knowledge of what nanostructure features and protein properties drive protein corona formation, both independently and in concert, it is possible to intelligently design nanostructures to bias the corona favorably. Harnessing the protein corona can enhance the efficiency of nanostructures *in vivo* by utilizing proteins to favorably improve circulation time, anatomical targeting, biocompatibility, cargo release efficacy, and cellular uptake. We anticipate this work will serve as a step toward the future of DNA constructs as nanomedicines, biosensors, and general tools for probing and manipulating biological organisms.

While we intend for this work to improve the engineering of DNA nanostructures for *in vivo* applications, we also acknowledge that several limitations and hurdles remain. While we maintained uniformity in experimental procedures and instrumentation, protein corona studies notoriously suffer from heterogeneity across laboratories.[76] For nanostructures, most envisioned applications *in vivo* are intended for nanostructure end-fate either on the cell membrane or within the cell. Therefore, while our study supports a better understanding of nanostructure physicochemical identity in human circulation, to thoroughly understand the role of the corona on nanostructure intracellular fate, this study bears repeating in other biological milieus like the cytoplasm. We expect results obtained using our enrichment classifier will be largely generalizable to other biological fluids, and other nanostructures, but this assumption needs to be experimentally verified. In addition, studies need to be performed with sequential incubation into different biologically relevant milieu. When a nanomaterial with preadsorbed protein corona enters circulation, certain *in vivo* proteins may adsorb and displace the original preadsorbed proteins as per the Vroman effect.[77] Nanostructures *in vivo* may traverse through numerous unique environments sequentially before reaching their intended target, and each of these environments drives the formation of protein coronas with unique identities. We expect this dynamic evolution of the corona would bias the corona's final composition to proteins with greater binding affinity to components of the nanostructure-corona complex, like nucleic acids and proteins. Furthermore, while this work focused on analyzing and predicting the protein corona of DNA nanostructures, further studies are needed to identify if other biomolecular constituents, such as lipids and metabolites, adsorb onto these nanostructures, and potentially impact their behavior and functionality. Lastly, while these are all factors that can be determined through further experiments and their integration with machine learning, there are inherent limitations to using machine learning algorithms altogether. Our algorithm provides insight into which proteins are likely to adsorb into the nanostructure protein corona, but it is generally unable to provide mechanistic insights into how and why. For such mechanistic and structural insights, further studies using high-resolution imaging and biochemical assays are necessary as a complement to the predictive power of machine learning. Taken together, we anticipate our study can enhance the effectiveness of DNA nanostructures *in vitro* and *in vivo*, and also inspire further approaches and inquiries into

https://doi.org/10.1021/acsnano.4c12259
ACS Nano XXXX, XXX, XXX−XXX

the important question of how, why, and which proteins adsorb to DNA nanostructures.

ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acsnano.4c12259.

Abbreviations used, materials used, experimental methods, computational methods, protein corona abundance plots, machine learning model performance metrics, characterizations of DNA nanostructures, DNA nanostructure design schematics, DNA nanostructure features, machine learning model features, DNA sequences (PDF)

AUTHOR INFORMATION

Corresponding Author

Grigory Tikhomirov − *Department of Electrical Engineering and Computer Sciences, University of California Berkeley, Berkeley, California 94720, United States;* orcid.org/0000-0001-6061-3843; Email: gt3@berkeley.edu

Authors

Jared Huzar − *Biophysics Graduate Group, University of California, Berkeley, Berkeley, California 94720, United States*

Roxana Coreas − *Department of Chemical and Biomolecular Engineering, University of California, Berkeley, California 94720, United States;* orcid.org/0000-0001-6057-4717

Markita P. Landry − *Department of Chemical and Biomolecular Engineering, University of California, Berkeley, California 94720, United States; Innovative Genomics Institute, Berkeley, California 94720, United States; California Institute for Quantitative Biosciences, University of California, Berkeley, Berkeley, California 94720, United States; Chan Zuckerberg Biohub, San Francisco, California 94158, United States;* orcid.org/0000-0002-5832-8522

Complete contact information is available at: https://pubs.acs.org/10.1021/acsnano.4c12259

Author Contributions

†J.H. and R.C. contributed equally to this work. J.H. synthesized, characterized, and purified DNA nanostructure library and developed machine learning models. R.C. performed experiments to obtain corona composition. J.H. wrote the manuscript with contributions from G.T., R.C., and M.P.L. M.P.L. and G.T. conceived the study. G.T. guided the project.

Notes

The authors declare no competing financial interest.

An earlier version of this manuscript is available on preprint server bioRxiv at Huzar J*, Coreas R*, Landry M, and Tikhomirov G. AI-based Prediction of Protein Corona Composition on DNA Nanostructures. 2024. bioRxiv. doi: 10.1101/2024.08.25.609594 (Posted August 26th, 2024).

ACKNOWLEDGMENTS

The authors thank Yifeng Shi, Myoungseok Kim, Durham Smith, and Arjun Banerjee for providing some of the DNA nanostructures and the UCSD Proteomics Facility for analyzing samples. J.H. was supported by NIH training grant (T32GM146614). G.T. was supported by Society of Hellman Fellows Fund, UC Berkeley, NSF CAREER (2240000) and NSF POSE Phase 2 (2346048) awards. We acknowledge the support of a Burroughs Wellcome Fund (BWF) Career Award at the Scientific Interface (CASI) (M.P.L.), a Dreyfus foundation award (M.P.L.), the Philomathia foundation (M.P.L.), an NSF CAREER award 2046159 (M.P.L.), an NSF CBET award 1733575 (to M.P.L.), a CZI imaging award (M.P.L.), a Sloan Foundation Award (M.P.L.), a McKnight Foundation award (M.P.L.), a Simons Foundation Award (M.P.L.), a Moore Foundation Award (M.P.L.), a Brain Foundation Award (M.P.L.), and a polymaths award from Schmidt Sciences, LLC (M.P.L.). M.P.L. is a Chan Zuckerberg Biohub investigator and a Helen Wills Neuroscience Institute Investigator. R.C. was supported by the NSF PRFB (2305663) and BWF PDEP awards. We acknowledge the use of BioRender for figure creation.

REFERENCES

(1) Rothemund, P. W. K. Folding DNA to create nanoscale shapes and patterns. *Nature* **2006**, *440*, 297−302.

(2) Seeman, N. C. Nucleic acid junctions and lattices. *J. Theor. Biol.* **1982**, *99*, 237−247.

(3) Mills, A.; Aissaoui, N.; Finkel, J.; Elezgaray, J.; Bellot, G. Mechanical DNA Origami to Investigate Biological Systems. *Adv. Biol.* **2023**, *7*, No. 2200224.

(4) Tikhomirov, G.; Petersen, P.; Qian, L. Fractal assembly of micrometre-scale DNA origami arrays with arbitrary patterns. *Nature* **2017**, *552*, 67−71.

(5) Wang, Z.; et al. DNA nanotechnology-facilitated ligand manipulation for targeted therapeutics and diagnostics. *J. Controlled Release* **2021**, *340*, 292−307.

(6) Kopperger, E.; et al. A self-assembled nanoscale robotic arm controlled by electric fields. *Science* **2018**, *359*, 296−301.

(7) Funke, J. J.; Dietz, H. Placing molecules with Bohr radius resolution using DNA origami. *Nat. Nanotechnol.* **2016**, *11*, 47−52.

(8) Wang, X.; et al. Collective cell behaviors manipulated by synthetic DNA nanostructures. *Fundam. Res.* **2023**, *3*, 809−812.

(9) Zhang, Y.; Ge, C.; Zhu, C.; Salaita, K. DNA-based digital tension probes reveal integrin forces during early cell adhesion. *Nat. Commun.* **2014**, *5*, 5167.

(10) Deng, Z.; et al. A light-controlled DNA nanothermometer for temperature sensing in the cellular membrane microenvironment. *Biosens. Bioelectron.* **2022**, *216*, No. 114627.

(11) Ohmichi, T.; et al. DNA-Based Biosensor for Monitoring pH in Vitro and in Living Cells. *Biochemistry* **2005**, *44*, 7125−7130.

(12) Wu, Z.; et al. Extraordinarily Stable Hairpin-Based Biosensors for Rapid Detection of DNA Ligases. *Biosensors* **2023**, *13*, 875.

(13) Douglas, S. M.; Bachelet, I.; Church, G. M. A Logic-Gated Nanorobot for Targeted Transport of Molecular Payloads. *Science* **2012**, *335*, 831−834.

(14) Ijäs, H.; Hakaste, I.; Shen, B.; Kostiainen, M. A.; Linko, V. Reconfigurable DNA Origami Nanocapsule for pH-Controlled Encapsulation and Display of Cargo. *ACS Nano* **2019**, *13*, 5959−5967.

(15) Meng, L.; et al. Tetrahedral DNA Nanostructure-Delivered DNAzyme for Gene Silencing to Suppress Cell Growth. *ACS Appl. Mater. Interfaces* **2019**, *11*, 6850−6857.

(16) Wang, Z.; et al. A Tubular DNA Nanodevice as a siRNA/Chemo-Drug Co-delivery Vehicle for Combined Cancer Therapy. *Angew. Chem., Int. Ed.* **2021**, *60*, 2594−2598.

(17) Kretzmann, J. A.; et al. Gene-encoding DNA origami for mammalian cell expression. *Nat. Commun.* **2023**, *14*, 1017.

(18) Xu, Z.; et al. Confinement in Dual-Chain-Locked DNA Origami Nanocages Programs Marker-Responsive Delivery of CRISPR/Cas9 Ribonucleoproteins. *J. Am. Chem. Soc.* **2023**, *145*, 26557−26568.

(19) Kim, M.; et al. Harnessing a paper-folding mechanism for reconfigurable DNA origami. *Nature* **2023**, *619*, 78−86.
(20) Kim, K.-R.; Kim, J.; Back, J. H.; Lee, J. E.; Ahn, D.-R. Cholesterol-Mediated Seeding of Protein Corona on DNA Nanostructures for Targeted Delivery of Oligonucleotide Therapeutics to Treat Liver Fibrosis. *ACS Nano* **2022**, *16*, 7331−7343.
(21) Li, J.; Fan, C.; Pei, H.; Shi, J.; Huang, Q. Smart Drug Delivery Nanocarriers with Self-Assembled DNA Nanostructures. *Adv. Mater.* **2013**, *25*, 4386−4396.
(22) Jiang, D.; et al. DNA origami nanostructures can exhibit preferential renal uptake and alleviate acute kidney injury. *Nat. Biomed. Eng.* **2018**, *2*, 865−877.
(23) Bastings, M. M. C.; et al. Modulation of the Cellular Uptake of DNA Origami through Control over Mass and Shape. *Nano Lett.* **2018**, *18*, 3557−3564.
(24) Chi, Q.; Yang, Z.; Xu, K.; Wang, C.; Liang, H. DNA Nanostructure as an Efficient Drug Delivery Platform for Immunotherapy. *Front. Pharmacol.* **2020**, *10*, 1585.
(25) Rodríguez-Franco, H. J.; Weiden, J.; Bastings, M. M. C. Stabilizing Polymer Coatings Alter the Protein Corona of DNA Origami and Can Be Engineered to Bias the Cellular Uptake. *ACS Polym. Au* **2023**, *3*, 344−353.
(26) Mahmoudi, M.; Landry, M. P.; Moore, A.; Coreas, R. The protein corona from nanomedicine to environmental science. *Nat. Rev. Mater.* **2023**, *8*, 422−438.
(27) Breznica, P.; Koliqi, R.; Daka, A. A review of the current understanding of nanoparticles protein corona composition. *Med. Pharm. Rep.* **2020**, *93*, 342−350.
(28) Barbero, F.; et al. Formation of the Protein Corona: The Interface between Nanoparticles and the Immune System. *Semin. Immunol.* **2017**, *34*, 52−60.
(29) Akhter, M. H.; et al. Impact of Protein Corona on the Biological Identity of Nanomedicine: Understanding the Fate of Nanomaterials in the Biological Milieu. *Biomedicines* **2021**, *9*, 1496.
(30) Mishra, R. K.; et al. Biological effects of formation of protein corona onto nanoparticles. *Int. J. Biol. Macromol.* **2021**, *175*, 1−18.
(31) Smolková, B.; et al. Protein Corona Inhibits Endosomal Escape of Functionalized DNA Nanostructures in Living Cells. *ACS Appl. Mater. Interfaces* **2021**, *13*, 46375−46390.
(32) Duan, Y.; et al. Prediction of protein corona on nanomaterials by machine learning using novel descriptors. *NanoImpact* **2020**, *17*, No. 100207.
(33) Ouassil, N.; Pinals, R. L.; Del Bonis-O'Donnell, J. T.; Wang, J. W.; Landry, M. P. Supervised learning model predicts protein adsorption to carbon nanotubes. *Sci. Adv.* **2022**, *8*, No. eabm0898.
(34) Ban, Z.; et al. Machine learning predicts the functional composition of the protein corona and the cellular recognition of nanoparticles. *Proc. Natl. Acad. Sci. U. S. A.* **2020**, *117*, 10492−10499.
(35) Baek, M.; et al. Accurate prediction of protein−nucleic acid complexes using RoseTTAFoldNA. *Nat. Methods* **2024**, *21*, 117.
(36) Ponnuswamy, N.; et al. Oligolysine-based coating protects DNA nanostructures from low-salt denaturation and nuclease degradation. *Nat. Commun.* **2017**, *8*, 15654.
(37) Langecker, M.; et al. Synthetic Lipid Membrane Channels Formed by Designed DNA Nanostructures. *Science* **2012**, *338*, 932−936.
(38) Charoenphol, P.; Bermudez, H. Aptamer-Targeted DNA Nanostructures for Therapeutic Delivery. *Mol. Pharmaceutics* **2014**, *11*, 1721−1725.
(39) Grossi, G.; Jepsen, M. D. E.; Kjems, J.; Andersen, E. S. Control of enzyme reactions by a reconfigurable DNA nanovault. *Nat. Commun.* **2017**, *8*, 992.
(40) Jungmann, R.; et al. Multiplexed 3D cellular super-resolution imaging with DNA-PAINT and Exchange-PAINT. *Nat. Methods* **2014**, *11*, 313−318.
(41) Lucas, C. R.; et al. DNA Origami Nanostructures Elicit Dose-Dependent Immunogenicity and Are Nontoxic up to High Doses In Vivo. *Small* **2022**, *18*, No. 2108063.
(42) Li, J.; et al. Self-Assembled Multivalent DNA Nanostructures for Noninvasive Intracellular Delivery of Immunostimulatory CpG Oligonucleotides. *ACS Nano* **2011**, *5*, 8783−8789.
(43) Ettig, R.; Kepper, N.; Stehr, R.; Wedemann, G.; Rippe, K. Dissecting DNA-Histone Interactions in the Nucleosome by Molecular Dynamics Simulations of DNA Unwrapping. *Biophys. J.* **2011**, *101*, 1999−2008.
(44) Marichal, L.; et al. From Protein Corona to Colloidal Self-Assembly: The Importance of Protein Size in Protein-Nanoparticle Interactions. *Langmuir* **2020**, *36*, 8218−8230.
(45) Francia, V.; Schiffelers, R. M.; Cullis, P. R.; Witzigmann, D. The Biomolecular Corona of Lipid Nanoparticles for Gene Therapy. *Bioconjugate Chem.* **2020**, *31*, 2046−2059.
(46) Putri, A. D.; et al. Differential cellular responses to FDA-approved nanomedicines: an exploration of albumin-based nanocarriers and liposomes in protein corona formation. *Nanoscale* **2023**, *15*, 17825−17838.
(47) Monopoli, M. P.; et al. Physical−Chemical Aspects of Protein Corona: Relevance to in Vitro and in Vivo Biological Impacts of Nanoparticles. *J. Am. Chem. Soc.* **2011**, *133*, 2525−2534.
(48) Casals, E.; Pfaller, T.; Duschl, A.; Oostingh, G. J.; Puntes, V. Time Evolution of the Nanoparticle Protein Corona. *ACS Nano* **2010**, *4*, 3623−3632.
(49) Resovi, A.; Pinessi, D.; Chiorino, G.; Taraboletti, G. Current understanding of the thrombospondin-1 interactome. *Matrix Biol.* **2014**, *37*, 83−91.
(50) Bouton, M.; et al. The under-appreciated world of the serpin family of serine proteinase inhibitors. *EMBO Mol. Med.* **2023**, *15*, No. e17144.
(51) Leung, L. L. K.; Morser, J. Carboxypeptidase B2 and carboxypeptidase N in the crosstalk between coagulation, thrombosis, inflammation, and innate immunity. *J. Thromb. Haemost.* **2018**, *16*, 1474−1486.
(52) Chekol Abebe, E.; et al. The structure, biosynthesis, and biological roles of fetuin-A: A review. *Front. Cell Dev. Biol.* **2022**, *10*, No. 945287.
(53) Cheung, M. C.; Vaisar, T.; Han, X.; Heinecke, J. W.; Albers, J. J. Phospholipid Transfer Protein in Human Plasma Associates with Proteins Linked to Immunity and Inflammation. *Biochemistry* **2010**, 7314.
(54) Meng, L.; et al. Effects of Lipopolysaccharide-Binding Protein (LBP) Single Nucleotide Polymorphism (SNP) in Infections, Inflammatory Diseases, Metabolic Disorders and Cancers. *Front. Immunol.* **2021**, *12*, No. 681810.
(55) Oskam, N.; et al. CD5L is a canonical component of circulatory IgM. *Proc. Natl. Acad. Sci. U. S. A.* **2023**, *120*, No. e2311265120.
(56) Lubbers, R.; van Essen, M. F.; van Kooten, C.; Trouw, L. A. Production of complement components by cells of the immune system. *Clin. Exp. Immunol.* **2017**, *188*, 183−194.
(57) Katsanou, V.; et al. HuR as a Negative Posttranscriptional Modulator in Inflammation. *Mol. Cell* **2005**, *19*, 777−789.
(58) Lord, M. S.; Melrose, J.; Day, A. J.; Whitelock, J. M. The Inter-α-Trypsin Inhibitor Family: Versatile Molecules in Biology and Pathology. *J. Histochem. Cytochem.* **2020**, 907.
(59) OJI, Y.; et al. The translation elongation factor eEF2 is a novel tumor-associated antigen overexpressed in various types of cancers. *Int. J. Oncol.* **2014**, *44*, 1461−1469.
(60) Schvartz, I.; Seger, D.; Shaltiel, S. Vitronectin. *Int. J. Biochem. Cell Biol.* **1999**, *31*, 539−544.
(61) Mariño-Ramírez, L.; Kann, M. G.; Shoemaker, B. A.; Landsman, D. Histone structure and nucleosome stability. *Expert Rev. Proteomics* **2005**, *2*, 719−729.
(62) Chyu, K.-Y.; et al. Immunization using ApoB-100 peptide−linked nanoparticles reduces atherosclerosis. *JCI Insight* **2022**, *7*, No. e149741.
(63) Georgila, K.; Vyrla, D.; Drakos, E. Apolipoprotein A-I (ApoA-I), Immunity. *Inflammation and Cancer. Cancers* **2019**, *11*, 1097.

https://doi.org/10.1021/acsnano.4c12259
ACS Nano XXXX, XXX, XXX−XXX

(64) Vihinen, M.; Torkkila, E.; Riikonen, P. Accuracy of protein flexibility predictions. *Proteins Struct. Funct. Bioinforma.* **1994**, *19*, 141−149.
(65) Song, Y.; et al. Discovery of Aptamers Targeting the Receptor-Binding Domain of the SARS-CoV-2 Spike Glycoprotein. *Anal. Chem.* **2020**, *92*, 9895−9900.
(66) Liu, Y.; et al. The Effects of Overhang Placement and Multivalency on Cell Labeling by DNA Origami. *Nanoscale* **2021**, *13*, 6819−6828.
(67) Lu, X.; et al. Providing Oligonucleotides with Steric Selectivity by Brush-Polymer-Assisted Compaction. *J. Am. Chem. Soc.* **2015**, *137*, 12466−12469.
(68) Chen, T.; Guestrin, C. XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 785−794 Association for Computing Machinery: New York, NY, USA, 2016, DOI: .
(69) Hajipour, M. J.; et al. An Overview of Nanoparticle Protein Corona Literature. *Small* **2023**, *19*, No. 2301838.
(70) Pedregosa, F.; et al. Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825−2830.
(71) Ho, T. K. Random decision forests. In *Proceedings of 3rd International Conference on Document Analysis and Recognition* IEEE: Montreal, QC, Canada, (1995). vol. *1* 278−282.
(72) Friedman, J. H. Greedy Function Approximation: A Gradient Boosting Machine. *Ann. Stat.* **2001**, *29*, 1189−1232.
(73) Lundberg, S.; Lee, S.-I. A Unified Approach to Interpreting Model Predictions. *arXiv* **2017**, Preprint at DOI: .
(74) Mohammad-Beigi, H.; et al. Mapping and identification of soft corona proteins at nanoparticles and their impact on cellular association. *Nat. Commun.* **2020**, *11*, 4535.
(75) Lacroix, A.; Edwardson, T. G. W.; Hancock, M. A.; Dore, M. D.; Sleiman, H. F. Development of DNA Nanostructures for High-Affinity Binding to Human Serum Albumin. *J. Am. Chem. Soc.* **2017**, *139*, 7355−7362.
(76) Ashkarran, A. A.; et al. Measurements of heterogeneity in proteomics analysis of the nanoparticle protein corona across core facilities. *Nat. Commun.* **2022**, *13*, 6610.
(77) Chen, F.; et al. Complement proteins bind to nanoparticle protein corona and undergo dynamic exchange in vivo. *Nat. Nanotechnol.* **2017**, *12*, 387−393.

https://doi.org/10.1021/acsnano.4c12259
ACS Nano XXXX, XXX, XXX−XXX